Report 5: Chronic Kidney Disease classification

Davron Khamidov, s273331,
ICT for Health attended in A.Y. 2021/22

December 30th, 2021

1 Introduction

Chronic kidney disease (CKD) derives from a gradual loss of kidney filtering capability over time, typically caused by high blood pressure and diabetes. Prevalence of the illness is around 10% in adult population, and its early detection avoids the dramatic consequence of complete kidney failure and necessity of kidney transplant.

Whilst a cure does not exist for CKD, treatments of kidney disease are available to reduce the symptoms, but they are expensive and impair the normal life of the affected subject (long dialysis sessions).

Kidney functionality can be assessed through the Glomerular Filtration Rate (GFR), calculated from the 24-hour collected urine or from the blood creatinine test.

A public dataset is available [1] to explore correlations between CKD and subject parameters. In particular, the dataset includes 24 features (see Table 1), among which 11 are numerical and 13 are categorical. Each of the 400 points of the dataset belongs either to class `ckd` (chronic kidney disease is present) or `notckd`. Unfortunately, some features are missing for some subjects (see Table 2) and must be replaced; on the contrary, there are no cases of missing class.

Object of the work is to use the dataset to build decision trees to classify new subjects as either healthy or affected by chronic kidney disease and measure the performance. Decision trees are all built using Python Scikit Learn class `DecisionTreeClassifier` [2] using entropy criterion; missing values are replaced using regression trees available in the same Python library [3].

2 Methods

2.1 Removal of rows with missing values

Table 2 shows that only 158 out of 400 rows have no missing values. If only these data are used, then most of the information is lost and the number of positive cases is 43, with a ratio $43/158 = 0.27$, which is much less than in the original dataset $250/400 = 0.62$. As a

	feature	meaning	type
1	age	age	numerical
2	bp	blood pressure (mm/Hg)	numerical
3	sg	specific gravity	categorical
4	al	albumin	categorical
5	su	sugar	categorical
6	rbc	red blood cells	categorical
7	pc	pus cell	categorical
8	pcc	ps cell clumps	categorical
9	ba	bacteria	categorical
10	bgr	blood glucose random (mg/dl)	numerical
11	bu	blood urea (mg/dl)	numerical
12	sc	serum creatinine (mg/dl)	numerical
13	sod	sodium (mEq/L)	numerical
14	pot	potassium (mEq/L)	numerical
15	hemo	hemoglobin (gms)	numerical
16	pcv	packet cell volume	numerical
17	wc	white blood cell count	numerical
18	rc	red blood cell count (million/cmm)	numerical
19	htn	hypertension	categorical
20	dm	diabetes mellitus	categorical
21	cad	coronary artery disease	categorical
22	appet	appetite	categorical
23	pe	pedal edema	categorical
24	ane	anemia	categorical

Table 1: Features in the UCI kidney dataset

consequence, the decision tree [2] based on just these 158 rows used as training dataset, shown in Figure 1, might be not completely correct. Notice that albumin ("al") is a categorical feature that takes values in the alphabet $\{0, 1, 2, 3, 4, 5\}$ where 0 means "normal" and 5 means "very abnormal/pathological" (i.e. very small quantities of albumin). It is therefore correct that a subject with categorical feature albumin less than 0.5 can be considered healthy. Note again that serum albumin levels less than 3.80 g/dL are associated with increased odds of rapid kidney function decline and increased risk of incident chronic kidney disease, but here feature "al" does not represent serum albumin quantities measured in g/dL, but degree of normality of the albumin quantity. However, among the 116 subjects with "al=0", there is just one subject affected by CKD, who is detected because of absence of hypertension ("htn" equal to zero). Of course this result cannot be generalized, and actually the software generates different decision trees each time it is run, since it can take other equivalent features to isolate the only subject positive to CKD. Therefore, the decision tree obtained from the reduced dataset only allows to find the importance of albumin in the diagnosis of CKD.

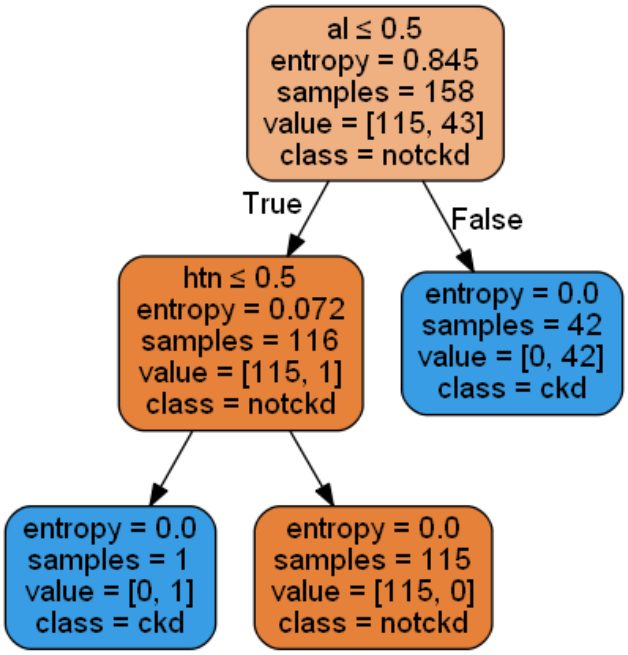

Figure 1: Decision tree obtained using only the rows without missing values

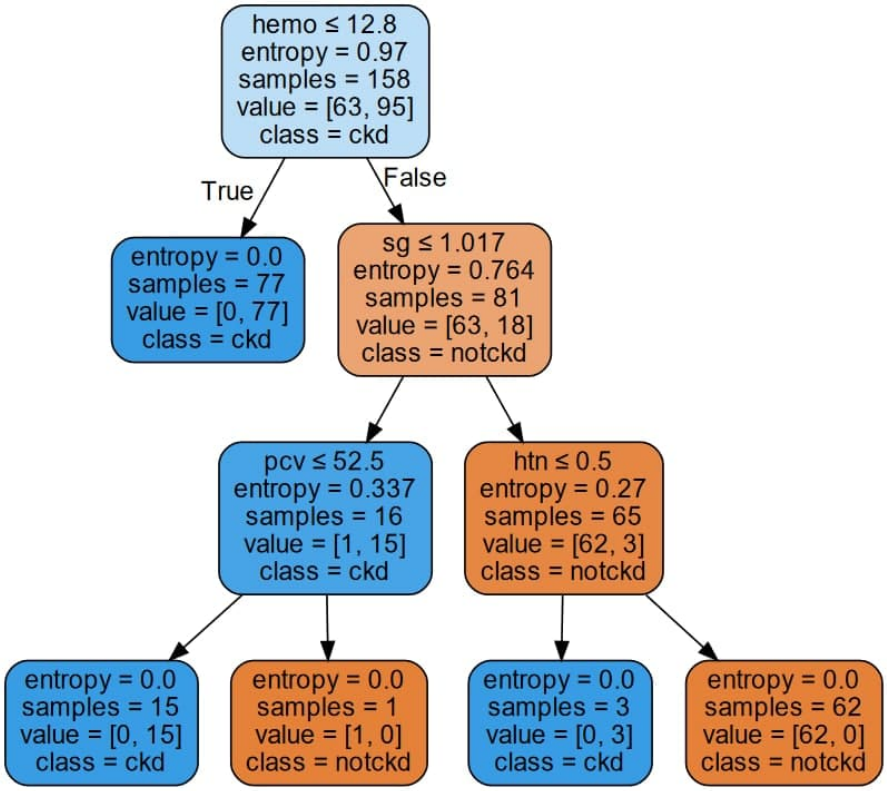

Figure 2: Decision tree obtained by replacing the missing values with the regressed values.

m	number of rows with m missing values
0	158
1	45
2	33
3	37
4	31
5	33
6	12
7	20
8	8
9	12
10	4

Table 2: Missing values in the dataset.

2.2 Substitution of missing with regressed values

The reduced dataset $\mathbf{Z}_{tr}$ with no missing values (described in Sect. 2.1) is used as training dataset to perform regression on the missing values. If only feature f is missing in row k, then the training regressor matrix $\mathbf{X}_{tr}$ is defined equal to $\mathbf{Z}_{tr}$ where column f is removed (158 rows and 23 columns), whereas the training regressand column $\mathbf{y}_{tr}$ is set equal to column f of $\mathbf{Z}_{tr}$. Matrix $\mathbf{Z}_{tr}$ and vector $\mathbf{y}_{tr}$ are used as inputs to train the tree regressor [3] and then the missing value in row k is substituted with the regressed value obtained by feeding the tree with the valid part of row k. If more than one feature is missing in row k, then exactly the same procedure is used, but the training regressand is a matrix instead of being a column.

Actually only the rows with up to 6 missing values (191) were included in this process, considering that regression accuracy cannot be sufficient if more than one fourth of the data is missing. Therefore, the obtained dataset after the replacement of the missing values is made of 349 rows, with 199 positive cases (ratio of positive cases 0.57, more similar to the ratio 0.62 of the original dataset). The new dataset is randomly shuffled and, to have a fair comparison with the result obtained in Sect. 2.1, 158 rows are used to train the decision tree. The obtained decision tree is shown in Fig. 2. Figure 2 shows the importance of the level of hemoglobin in the diagnosis of CKD. The medical condition in which the hemoglobin is less than normal causes anemia. Anemia is a common complication of chronic kidney disease. In men, anemia is typically defined as a hemoglobin level of less than 13.5 gram/100 ml and in women as hemoglobin of less than 12.0 gram/100 ml. The decision tree obtained from the shuffled data with seed 1 shows that level of hemoglobin below the 12.8 plays a major role in the diagnosis of CKD.

Knowing that decision trees tend to overfit, shuffling was performed 3 times and 3 different decision trees were obtained. According to other decision trees, when diagnosing CKD,

attention should be paid to also an increase in serum creatinine and a decrease in packet cell volume.

3 Accuracy, sensitivity, specificity

The decision tree of Sect. 2.1, obtained with the reduced dataset of 158 points, was used to classify the 191 points of the dataset with missing values regressed as described in Sect. 2.2. The decision trees obtained in Sect.2.2 were used to classify the 191 points not belonging to training dataset.

Accuracy, sensitivity and specificity were measured several times, using different state seeds in the generation of the decision tree [2], and several shuffes for the decision trees of Sect. 2.2. Results are given in table 3. In the first case, when the data were not shuffled despite high sensitivity, the accuracy was satisfied. In the following cases, when the data was shuffled with different seeds, the sensitivity decreased slightly and the accuracy increased significantly. All these cases showed approximately the same accuracy, sensitivity and specificity. When the random forest algorithm was run instead of the decision tree algorithm, the accuracy became slightly better

	No shuffle	Seed 1	Seed 2	Seed 3	Random Forest
Accuracy	0.853	0.989	0.978	0.978	0.994
Sensitivity	0.675	0.539	0.581	0.565	0.539
Specificity	0.263	0.445	0.392	0.408	0.45

Table 3: Comparison among the

4 Conclusions

Despite the good results obtained on the random forest algorithm, decision tree algorithms give more important features. Using this algorithm, it was found that in the diagnosis of CKD, such features as the level of hemoglobin, albumin, serum creatinine, and packet cell volume are important. On the other hand, the random forest algorithm, despite good statistical results, did not give any features for diagnosing CKD

References

[1] `https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease`

[2] `https://scikit-learn.org/stable/modules/tree.html#classification`

[3] `https://scikit-learn.org/stable/modules/tree.html#regression`